Current Statistical Methods for Spatial Epidemiology: A Review

Osei FB
Department of Mathematics and Statistics, University of Energy and Natural Resources, Ghana

*Corresponding author: Osei FB, Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana

Received: October 13, 2014; Accepted: November 11, 2014; Published: November 25, 2014

Introduction

Spatial epidemiology is the study of the spatial/geographical distribution of disease incidences and its relationship to potential risk factors. Knowledge of the spatial variations of diseases and characterization of its spatial structure is essential for the epidemiologist to better understand the population’s interaction with its environment. The origin of spatial epidemiology dates back to 1855 with the classic epidemiologic studies of John Snow on cholera transmission. Snow’s study of London’s cholera epidemic provides one of the most famous examples of spatial epidemiology. Mapping the locations of cholera victims, Snow was able to trace the cause of the disease to a contaminated water source. Spatial analysis in the nineteenth and twentieth century mostly took the form of plotting the observed disease cases or rates [1]. Advances in technology now allow not only disease mapping but also the application of spatial statistical methods, such as cluster analysis [2,3] and ecological analysis [4-6] in epidemiological research. Geographic Information System (GIS) methods and modern statistical methods allow an integrated approach to address both tasks; i.e. inference on the geographical distribution of the disease and its prediction at new locations. Many diseases are influenced by environmental variables, and since these variables are spatially continuous in nature, the disease rates tend to exhibit spatial dependency, popularly known as spatial autocorrelation. Thus such patterns of spatial autocorrelation confirm the natural law of nature, popularized by Tobler [7] as the first law of geography: “Everything is related to everything else, but near things are more related than distant things”. The use of standard/classical statistical techniques for modeling spatially distributed diseases either leads to over estimation or under estimation of parameters in question. The objective of this manuscript is to provide a review of the current statistical methods that are useful in analyzing and modeling spatially distributed diseases, their relative weaknesses and strength. Much attention and focus is provided for methods which are relatively advantageous and widely used.

Cluster Analysis

Fundamental to the spatial epidemiologist is the investigation of possible disease clustering. Cluster analysis provides opportunities for the epidemiologist to understand the spatial distribution of diseases and the possible association between demographic and environmental exposures [8-11]. Searching for disease clustering involves an assessment of local or global accumulation of the disease incidences [12,13]. The focus of global cluster analysis is to determine the presence or absence of clustering in the whole study region. There are numerous methods for testing global clustering, including those proposed by Alt and Vach [14], Besag and Newell [8], Cuzick and Edwards [15], Diggle and Chetwynd [16], Grimson [17], Moran [18], Tango [19-21], Walter [22-24] and Whittemore et al. [25]. The most widely used measure of global clustering in epidemiology is the method proposed by Moran [18]. Moran’s Index is a weighted correlation coefficient that is used to measure deviation from spatial randomness. The Index $I_{m}$ statistic is similar to the Pearson correlation coefficient [18,26,27] with the form:

$$I_{m} = \frac{\sum\sum w_{i} (r_{i} - \bar{r})(r_{j} - \bar{r})}{\sum\sum (r_{i} - \bar{r})^{2}}$$

(1)

where $N$ is the number of spatial objects, $w_{i}$ is the element in the spatial weights matrix corresponding to the spatial object pairs $i, j$; and $r_{i}$ and $r_{j}$ are the observed rates for objects $i$ and $j$ with mean rate $\bar{r}$. When the weights are not row-standardized, the scaling factor $N/\sum$ is applied, such that $S_{m} = \sum w_{i}$. Values range from −1 (indicating perfect dispersion) to +1 (perfect clustering or deviation from randomness). Negative (positive) values indicate negative (positive) spatial autocorrelation.

Deviation from spatial randomness indicates specific spatial arrangements of geographic location information such as clusters [18]. Although Moran’s Index was originally developed to analyze continuous data, its application to analyze count data of health events
Global cluster analysis can obscure local effects since the assumption of stationary is rarely met. Local cluster analysis defines the characteristics of the clusters, such as size, location and intensity. Several formal methods and techniques for identifying local disease clusters have been developed for both point and areal level data [8,9]. Examples of local clustering methods include spatial correlograms [35-39] the Local Indicator of Spatial Association [40], the local $G^*_i$ statistics [41], Ripley’s $K$-function [42-44], Cluster Evaluation Permutation Procedure (CEPP) [45], the Knox test [46,47], and Kulldorff’s spatial scan statistic [2]. Other methods for space-time clustering include Mantel’s test [48], Ederer-Meyer-Mantel test [49], Barton’s test [50], Diggle et al. test [51], Jacquez’s $k$ nearest neighbors test, and Kulldorff’s space-time scan statistic [2].

The spatial scan statistic developed by Kulldorff [10,11,52] offers several advantages over the others: (1) it corrects for multiple comparisons, (2) it adjusts for the heterogeneous population densities among the different areas in the study, (3) it detects and identifies the location of the clusters without prior specification of their suspected location or size thereby overcoming pre-selection bias, (4) and allows adjustment for covariates. Also Kulldorff’s spatial scan statistic is both deterministic (i.e., it identifies the locations of clustering) and inferential (i.e., it allows for hypothesis testing and evaluation of significance). The spatial scan statistic has been used to detect and evaluate various disease clusters including leukemia [9,53], cancer [10,45,53-56], giardiasis [57], tuberculosis [58], diabetes [59], Creutzfeldt-Jacob disease [60], granulocytic ehrlichiosis [61], and amyotrophic lateral sclerosis [62].

The flexible spatial scan statistic is a recent cluster detection methodology developed by Takahashi and Tango. This approach is based on the original idea of Kulldorff. Unlike Kulldorff’s approach, however, which imposes a circular window to define the potential cluster areas [9], Takahashi and Tango’s flexible spatial scan statistic imposes an irregularly shaped window on each region connecting its adjacent regions.

For any given location $i$, a set of irregularly shaped windows consisting of $k$ connected locations including $i$ moves from 1 to a pre-set maximum window size $K$ (which is proportional to the population at risk). To avoid detecting a cluster of an unlikely peculiar shape, the connected locations are restricted as the subsets of the set of location $i$ and $(K - 1)$-nearest neighbors to location $i$. In effect a very large number of different but overlapping arbitrarily shaped windows are created. For location $i$, the flexible scan statistic considers $K$ concentric circles plus all the sets of connected locations, including location $i$, whose centroids are located within the $K$-th largest concentric circle. Let $W_{i;k} = j_1,…,j_k$ denote the $k$-th window which is a set of $k$ regions connected starting from the region $i$, where $j_k$ is the number of $j$ satisfying $W_{i;k} \subseteq W_{i;j}$ for $k = 1,…,K$. Then, all the windows to be scanned are included in the set:

$$W = \{w_{ik}; 1 \leq i \leq m, 1 \leq k \leq K, 1 \leq j \leq j_k \}. \tag{2}$$

Under the alternate hypothesis, there is at least one window $W$ for which the underlying risk is higher inside the window when compared with outside. For each window the likelihood of the observed number of occurrences within and outside the window under the Poisson assumption is computed as:

$$I(W) = \sup_{\theta \in W} \left( \frac{E(W)}{E(W)} \right) \left( \frac{E(O)}{E(O)} \right) \left( \frac{E(O)}{E(O)} \right). \tag{3}$$

Where $W$ indicates all the regions outside the window $W$, and $O(\cdot)$ and $E(\cdot)$ denote the observed and expected number of occurrences within the specified window, respectively. The indicator function $I(\cdot)$ is 1 when the number of occurrences within the window is more than the expected number and 0 otherwise. The window $W^*$ that attains the maximum likelihood is defined as the Most Likely Cluster (MLC). This approach is able to detect arbitrarily shaped clusters, and this statistic is well suited for detecting and monitoring disease outbreaks in irregularly shaped areas.

Popular software packages for conducting cluster analysis includes Sat Scan for circular spatial scan statistics developed by Martin Kulldorff [11] and FleX Scan developed by Tango and Takahashi [63] for flexible shaped scan statistics. Sat Scan can implement both purely spatial and space-time cluster analysis; however, these are not yet implemented in FleX Scan. The scan statistics technique has also been implemented the SpatialEpi [64] package of the R software for statistical computing.

**Ecological Analysis**

A significant interest in spatial epidemiology lies in identifying associated risk factors which enhance the risk of infection, the so called ecological analysis [65,66] or geographic correlations studies [67]. The term ecological analysis is used loosely here to denote associating aggregated disease outcomes with related risk factors or covariates, where inference still remains at the aggregated level.

**Classical linear methods**

The most prominent method is the classical linear regression model, where the response variable $y$ is assumed to be independent normal or Gaussian distributed and covariates, say $x_1,…,x_p$ act linearly on the response. By assumption, the conditional expectation of $y$ is:

$$\eta = E(y|x_1,...,x_p) = \sum \beta_j x_j + \epsilon,$$

where the regression coefficients $\beta_1,…,\beta_p$ determine the strength of the influences of the covariates, and the linear predictor $\eta$ is the sum of the covariate effects. Here, each observation has an underlying mean of $\Sigma \chi_j \beta_i$ and normally distributed random error term $\epsilon$. Generally, the random error term $\epsilon = (\epsilon_1,\epsilon_2,…,\epsilon_p)$ has zero mean and uncorrelated variance-covariance matrix $\Sigma_{\epsilon}$, i.e. $E(\epsilon) = 0$ and $\Sigma_{\epsilon} = \text{Var}(y) = \text{Var}(\epsilon) = \sigma^2I$, and $I$ is a $p \times p$ identity matrix. The assumption of independent observations also implies that $E(\epsilon \epsilon^T) = E(\epsilon) E(\epsilon) = 0$.

For disease counts of small areas with relatively small populations at risk and few observed cases, rates may not follow the assumptions of the linear model. In such cases, a direct connection between the expectation of $y$ and the linear predictor $\eta$ is not possible. Generalized Linear Models (GLMs) extend the classical linear model for Gaussian responses to more general situations such as binary or count data [68-71] to ensure the appropriate domain of $E(y|x_1,...,x_p)$. By introducing
a more general transformation or response function \( h \), equation (1.1) can be rewritten as:

\[
\eta = E(y|x_1,\ldots,x_p) = h(\beta_0 + x_1\beta_1 + \ldots + x_p\beta_p).
\]

(5)

Both the classical linear model and GLMs provide the means to quantify and describe only first-order effects or small-scale variation in the mean of the disease outcome. These methods ignore second-order spatial effects or small-scale variations that arise from interactions between neighbors, i.e., spatial autocorrelation. Both methods assume that any spatial pattern observed in the outcome \( y \) is entirely due to the spatial patterns in the covariates; therefore, no residual spatial variation is accounted for. If an important covariate is inadvertently omitted, however, estimates of \( \beta \) will be biased [72], and if this covariate varies spatially, residual spatial variation will often manifest itself as spatial autocorrelation in the residual process. Hence when these methods are used to analyze spatially correlated data, the standard error of the covariate parameters would be underestimated and thus the statistical significance would be overestimated [73].

**Spatial methods**

Spatial statistical methods, such as spatial regression, incorporate spatial autocorrelation according to the way spatial neighbors are defined. A spatial regression model may be parameterized as equation (4). A modification of the variance-covariance matrix \( \Sigma \) is then required to allow spatially correlated error terms. Common methods to incorporate spatially correlated error terms in the variance-covariance matrix \( \Sigma \) is the Simultaneous spatial Autoregressive (SAR), Conditional spatial Autoregressive models (CAR), and Spatial Moving Average models (SMA). Both the SAR and CAR correspond to autoregressive procedures in time series analysis [43]. These models are well explained in Cliff and Ord [26], Haining [74], Ripley [43], and Cressie [73].

Under CAR model specification, the conditional expectation of the response variable \( y \) is specified as

\[
\eta = E(y|x_1,\ldots,x_p) = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p + pw[y-(\beta_0 + x_1\beta_1 + \ldots + x_p\beta_p)] + \epsilon,
\]

which can be surmised as

\[
\eta = E(y|x_1,\ldots,x_p) = \Sigma \beta + pw(y-\Sigma \beta) + \epsilon,
\]

and simplified in matrix notation as \( Y = X\beta + \rho W(Y-X\beta) + \epsilon \). The error terms assumed normally distributed with zero mean and therefore not suitable for modeling directional processes. Also, both methods assume that any spatial pattern observed in the outcome \( y \) is entirely due to the spatial patterns in the covariates; therefore, no residual spatial variation is accounted for. If an important covariate is inadvertently omitted, however, estimates of \( \beta \) will be biased [72], and if this covariate varies spatially, residual spatial variation will often manifest itself as spatial autocorrelation in the residual process. Hence when these methods are used to analyze spatially correlated data, the standard error of the covariate parameters would be underestimated and thus the statistical significance would be overestimated [73].

**Generalized structured additive regression**

Generalized Additive Models (GAM) also provides a powerful class of models for modeling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. Modeling the nonlinear effects of continuous covariates may be based on smoothing splines [77], local polynomials [78], regression splines with adaptive knot selection [79-81] and P-splines [82,83].

For a spatial lagged model, spatial autocorrelation is included as an additional predictor in the form of spatially lagged dependent variable. Thus \( Y = \rho Y^* + X\beta + \epsilon \), where the lagged dependent variable is \( Y = WY \), which finally yields

\[
Y = (1-\rho Y)^*X\beta + (1- \rho Y)^* \epsilon.
\]

(8)

Where it is believed that the autoregressive process occur only in the error terms rather than either in the in the response or in the predictor, the OLS model \( Y = X\beta + \epsilon \) is complemented by a spatially lagged error term of the form \( \epsilon = \lambda W \epsilon + \epsilon \). This yield

\[
Y = \lambda W \epsilon + X\beta + \epsilon,
\]

(9)

Where \( \lambda \) is the lagged-error variable.

Where it is believed that spatial autocorrelation affects both the response and predictor variables, then another term \( WY \) which expresses a lagged-decency of the predictor variables is added to the model. This results in a spatial lagged mixed model of the form:

\[
Y = \lambda W \epsilon + X\beta + WY + \epsilon,
\]

(10)

Where \( \gamma \) expresses the regression coefficient of the lagged-error variable.

Haining [74] expresses the facts that ever SAR model is also a CAR model with \( K = S + S^T \cdot S \), where \( K \) is the \( \rho W \) of the CAR model and \( K \) is the \( \rho W \) of the SAR model.

Numerous software packages have been developed for implementing spatial regression models. Typical amongst them is the free software GeoDa [75] which easily fits both spatial lag and error models. The sped package in R software has vigorous functions for fitting spatial regression models. The comprehensive econometric toolbox developed by LeSage and Pace [76] in MATLAB has numerous functions for fitting spatial regression models.

CAR models restrict the spatial weight matrix to be symmetrical and therefore not suitable for modeling directional processes. Also, the \( k \) nearest neighbor connectivity option for \( w \) generates as asymmetric neighborhood structure and therefore not suitable for CAR models.

SAR model on the other hand can be specified under three different variants. As spatial lagged model, as spatial error model or spatial lagged mixed-model. Unlike CAR models, the neighborhood connectivity matrix \( W \) in the SAR model need not be symmetrical.

For a spatial lagged model, spatial autocorrelation is included as an additional predictor in the form of spatially lagged dependent variable. Thus \( Y = \rho Y^* + X\beta + \epsilon \), where the lagged dependent variable is \( Y = WY \), which finally yields

\[
Y = (1-\rho Y)^*X\beta + (1- \rho Y)^* \epsilon.
\]

(8)

Where it is believed that the autoregressive process occur only in the error terms rather than either in the in the response or in the predictor, the OLS model \( Y = X\beta + \epsilon \) is complemented by a spatially lagged error term of the form \( \epsilon = \lambda W \epsilon + \epsilon \). This yield

\[
Y = \lambda W \epsilon + X\beta + \epsilon,
\]

(9)

Where \( \lambda \) is the lagged-error variable.

Where it is believed that spatial autocorrelation affects both the response and predictor variables, then another term \( WY \) which expresses a lagged-decency of the predictor variables is added to the model. This results in a spatial lagged mixed model of the form:

\[
Y = \lambda W \epsilon + X\beta + WY + \epsilon,
\]

(10)

Where \( \gamma \) expresses the regression coefficient of the lagged-error variable.

Haining [74] expresses the facts that ever SAR model is also a CAR model with \( K = S + S^T \cdot S \), where \( K \) is the \( \rho W \) of the CAR model and \( K \) is the \( \rho W \) of the SAR model.

Numerous software packages have been developed for implementing spatial regression models. Typical amongst them is the free software GeoDa [75] which easily fits both spatial lag and error models. The sped package in R software has vigorous functions for fitting spatial regression models. The comprehensive econometric toolbox developed by LeSage and Pace [76] in MATLAB has numerous functions for fitting spatial regression models.

**Generalized structured additive regression**

Generalized Additive Models (GAM) also provides a powerful class of models for modeling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. Modeling the nonlinear effects of continuous covariates may be based on smoothing splines [77], local polynomials [78], regression splines with adaptive knot selection [79-81] and P-splines [82,83].
covariate $u_i$. The linear combination $u_i'y$ corresponds to the usual parametric part of the predictor. The function $f_{out} (S)$ accounts for spatial effects of the data.

**Bayesian Estimation**

STAR models are highly parameterized; therefore, inference is based on a fully Bayesian estimation of the posterior distribution of the model parameters rather than maximum likelihood estimation methods. Since the posterior is analytically intractable, the parameter estimates are generated by drawing random samples from the posterior via MCMC simulation techniques.

Bayesian estimation and inference in statistical modeling provides a number of advantages over the classical approaches. This includes a more natural interpretation of parameter intervals, and the ease with which the true parameter density may be obtained. Bayesian approach has recently been given intense focus due to the widespread adoption of Markov Chain Monte Carlo (MCMC) methods. In the past, Bayesian estimation and inference was often daunting due to the requirement of numerical integration. The MCMC estimation method decomposes complicated estimation problems into simpler problems that rely on conditional distributions for each parameter in the model [87]. In classical approaches such as maximum likelihood estimation, inference is based on the likelihood of the data alone. In Bayesian approach, the likelihood of the observed data $y$ given a $d$ dimensional parameter set $\theta = (\theta_1, \ldots, \theta_d)$, denoted as $p(y|\theta)$, is used to modify the prior beliefs $p(\theta)$ with the updated knowledge summarized in a posterior density $p(y|\theta)$. Applying Bayes theorem, $p(\theta|y) = p(y|\theta)p(\theta)/p(y)$ is found, where the marginal likelihood $p(y)$ is obtained by integrating the likelihood over the prior densities, i.e. $p(y) = \int p(y|\theta)p(\theta)d\theta$. Since $p(y)$ can be regarded as a normalizing constant, the posterior density can be simplified as $p(\theta|y) = p(y|\theta)p(\theta)$.

**Priors for unknown functions and fixed effects**

The unknown functions $f_1, \ldots, f_p$ and the fixed effects $\beta$ are considered as random variables and must be supplemented by appropriate prior assumptions. In the absence of any prior knowledge, diffuse prior $p(\beta)$ can be assigned in the fixed effects. Alternatively, a weak informative multivariate Gaussian distribution may be assigned. For modeling the unknown functions $f_1, \ldots, f_p$, there exists a variety of different approaches. Polynomials of degree $l$ are often not flexible enough for small $l$, yet estimates become more flexible but also rather unstable for large $l$, especially at the boundaries [85]. Eilers and Marx [82] suggest specific forms of polynomial regression splines which are parameterized in terms of B-spline basis functions together with a penalization of adjacent parameters, also known as P-splines. For instance, following Eilers and Marx [82], $f(x)$ can be approximated by a polynomial spline of degree $l$ with equally spaced knots $x_{min} = \xi_0 < \xi_1 < \cdots < \xi_{l-1} < \xi_{max}$ within the domain of $x$. The assumption that $f(x)$ can be approximated by a polynomial spline leads to a representation in terms of a linear combination of $d=s+l$ basic functions $R_s$, i.e. $f(x) = \sum_{s=0}^{s+l} \beta_s R_s(x)$. Thus, the estimation of $f(x)$ is reduced to the estimation of the vector of unknown regression coefficients $\beta = (\beta_0, \ldots, \beta_{s+l})$ from the data. Detailed description of Bayesian P-Splines in STAR models can be found in Brezger [85].

**Priors for spatial effects**

The spatial effect is commonly introduced in a hierarchical fashion via prior distributions of location-specific random effects. Unlike the SAR, CAR, or SMA models, spatial dependencies are estimated for each spatial unit. A major significance of STAR modeling approach is that the spatial effect can be split into spatially structured (correlated) and a spatially unstructured (uncorrelated) effects. Thus, $f_{out}(s) = f_{str}(s) + f_{unstr}(s)$ where the function $f_{str}(s)$ accounts for spatially correlated effects of the data, whereas the function $f_{unstr}(s)$ accounts for unobserved heterogeneity, occurring locally or at a large scale. The most common prior for modeling the structured spatial effects $f_{str}(s)$ is the Markov random field prior pioneered by Besag [88,89]:

$$p(f_{str}(s)|f_{str}(s'), s' \neq s, \gamma^2) \sim N \left( \frac{\sum f_{str}(s'), \gamma^2}{\gamma^2} \right).$$  (12)

Here $s = (1, \ldots, S)$ represents the locations of connected geographical regions, $N$ is the number of geographical neighbors and $s' \neq s$ denotes that geographical locations $s'$ and $s$ are neighbors. The uncorrelated $f_{unstr}(s)$ part may be estimated based on location-specific Gaussian random effects $p(f_{unstr}(s) | r^2_{unstr}) \sim N(0, r^2_{unstr})$. In a fully Bayesian estimation, hyper-priors for the variance parameters $r^2_j = j=1, \ldots, p$, $r^2_{unstr}$ are also considered as unknown; therefore, appropriate hyper-parameters have to be assigned. Commonly, highly dispersed, but proper, inverse Gamma distributions $p(r^2) \sim IG(a, b)$ with known hyper-parameters $a$ and $b$ with density function $p(r^2) \propto r^2^{a-1}e^{-br^2}/r^2$ are assigned in the second stage of the hierarchy.

Different forms of STAR models may be structured for both cross-sectional and longitudinal data. Well known models that can be structured include GAM, Generalized Additive Mixed Models (GAMM), spatial regression models, generalized geodiffusive mixed models (GGAMM), dynamic models, varying coefficient models, and geographically weighted regression [90] may be useful within a unifying framework. Detailed description of these models and their applications can be found in Fahrmeir and Lang [91,92], Lang and Brezger [93], Brezger and Lang [94], Eilers and Marx [82], Marx and Eilers [83], Wahba [95], and Hastie and Tibshirani [77].

Much literature has been developed around methodological issues relating to the Bayesian approach [96-103]. Bayesian approaches to GAM are currently either based on regression splines with adaptive knot selection [104-110], or on smoothness priors [77,91,92]. The development and implementation of Markov Chain Monte Carlo (MCMC) methods in software such as WinBUGS [111] and BayesX [112] have made Bayesian estimation approaches simpler.

**Conclusion**

Space has become, and would continue to be, an essential dimension in epidemiology. This is mainly due to the availability and quality of geographically referenced health data. Thus the relevance of space is unlimited, both in theory and in practice. However, statistical methods for spatial epidemiologic data are limited in the mainstream statistics literature. Many texts in the field of spatial statistics and related fields address the significance of space and theoretical approaches in diverse forms. This manuscript has discussed a wide range of statistical methods useful in spatial epidemiology; focusing on those relevant under two main themes; cluster analysis and ecological analysis. The availability of open source software packages designed to facilitate such methods and techniques has been key and resourceful in their implementation. However, their implementations must be guided by good practice theories within the epidemiologic
principles. With these, spatial epidemiologic studies will continue to play a critical key role in the understanding of disease epidemiology, especially the complex relationship between population, health and environment.

References


